PAIRWISE CLASSIFICATION BASED ON MORPHOLOGICAL AND SEMANTIC FEATURES FOR COREFERENCE RESOLUTION IN RUSSIAN

Serebrennikova Yulia, RSUH, Moscow (julie.mamaeva@yandex.ru)
Budnikov Egor, ABBYY, Moscow (egor.budnikov@abbyy.com)

Coreference resolution is the task of determining which elements of a text refer to the same entity. To solve this task means to identify all coreferent elements and attribute them to the correct cluster. Elements referring to the same entity must be attributed to the same cluster, while elements referring to different entities must be attributed to different clusters.

In this paper, we use a pairwise binary classification model as an approach to coreference resolution task for the Russian language, which classifies all pairs of mentions of an entity in the text (m_1, m_n) as coreferent or non-coreferent. As train and test data, the OpenCorpora corpus was used. This corpus of texts in Russian has morphological, syntactical and semantic layers and is freely and fully accessible to researchers.

The contribution of the present research is the analysis of morphological and semantic features' impact on the accuracy of the prediction model. To determine this impact, several experiments were performed. As a learning method for the classification model, random forest ensemble was used. The accuracy of the model is estimated after each experiment by the values of the existing coreference resolution evaluation metrics (MUC-6, B-CUBE and CEAFE). Each metric value is an average of values of the same metric computed for each text in the test corpora. Precision, recall and F-measure are computed for each metric. Four experiments were performed. Errors were analyzed after performing each experiment. Building on the error analysis, relevant morphological and semantic features were added. The last experiment was aimed at enhancing the accuracy of the model by combining different sets of morphological features and adding some new features to the model. At the end of the article, we discuss the importance of individual features, draw some conclusions based on the results of the classification model and propose how to raise the coreference chains accuracy.

Keywords: coreference resolution, scoring metrics, machine learning, XGBoost, gradient boosting

1. Introduction

Coreference resolution is the task of determining which elements of a text refer to the same entity. Textual phrases that refer to real-world objects or events are called mentions. Entities are real-world objects or events. Two mentions referring to the same entity are called coreferent mentions. Sometimes the right mention is called an anaphor, and the left mention an antecedent [3]. We adhere to this terminology in this paper.

Example: A boy$_1$ has lost$_2$ his$_1$ wallet$_3$. He$_1$ can't find it$_3$ anywhere. He$_1$ will be late because of it$_{2.}$

To solve coreference resolution task means to identify all coreferent elements and attribute them to the correct cluster. Elements referring to the same entity must be attributed to the same cluster, while elements referring to different entities must be attributed to different clusters.

Most of the researches were targeted coreference resolution for texts in the English language [7]. The Russian language has more morphological richness than English does, hence coreference resolution for Russian is a more complicated task.

In this paper, we use a pairwise binary classification model based on morphological features as an approach to coreference resolution task for Russian, which classifies all pairs of mentions of an entity in the text (m_1, m_n) as coreferent or non-coreferent. Our main objective is to test different morphological and semantic features and their impact on model accuracy. In section 2 we describe existing methods to coreference resolution task. Common architecture description of our model and our approach for reassembling chains from classifier output are detailed in Section 2. The corpus that is used as training and test dataset and its peculiarities are describe in Section 4. In Section 5, we describe the evaluation measures. Our experiments are discussed in Section 6.

2. Existing methods

Most of the approaches to the problem of coreference resolution are treated as the problem of classification [6]. These approaches can be based on mention-pair model or entity-mention model. The essence of first one is to divide all the mentions in the texts in pairs and to determine whether two mentions are coreferent or not. Compared with the mention-pair counterpart, the entity-mention model aims to make coreference decision at an entity level. Classification is done to determine whether a mention is a referent of a partially found entity [2] i.e. entity-mention models are trained to determine whether an active mention belongs to a preceding, possibly partially-formed, coreference cluster. Hence, they can employ cluster-level features (i.e., features that are defined over any subset of mentions in a preceding cluster) [8].

The task of coreference resolution can be treated also as the problem of ranking. Supporters of this approach consider ranking to be a more natural reformulation of coreference resolution than classification, as a ranker allows all candidate antecedents to be considered simultaneously and therefore directly captures the competition among them [8]. Methods based on ranking of mentions use similar approach to constructions of sets of mentions [8]. The ranking model is trained to answer which of preceding mentions is the likeliest antecedent. To do so, the ranking model sorts all pairs of mentions that include a given anaphor and a preceding mention in the order of a decrease of probability that they are coreferent.

3. Our approach and common architecture description

The task of coreference resolution is considered as mention-pair binary classification as the most common approach. Within this scenario objects are pairs of mentions mi∊d of a document. Set of pairs S contains all pairs (m_i ,m_j) , i< j . Evaluation score is between 0 and 1. All pairs that have a score more than 0.5 will have a '1' label. As an algorithm of supervised-machine learning algorithm gradient boosting was used. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. As implementation of Gradient boosting framework XGBoost was used. This optimized distributed gradient boosting library provides a parallel tree boosting. XGBoost uses decision tree ensembles as the model choice. A tree ensemble is much stronger than a single tree as the prediction scores of each individual tree in a tree ensemble are summed up to get the final score, so each tree tries to complement each other [4]. Since we aimed to perform several experiments to compare different sets of features, our choice fell on this learning algorithm due to quite high learning speed of a model based on it. Moreover, XGBoost library contains some tools for comparing feature importances.

To obtain a set of chains within a document, mentions that were classified as coreferent by the classifier were reassembled in a chain. To accomplish it, we built an undirected graph to represent coreferent mentions (m_i ,$m_{j)}$ where vertices are mentions. There is an edge e_i between vertices v_i and v_j if the mention m_i represented by v_i is coreferent to mention m_j represented by $v_{j.}$ Depth-first search algorithm was used to reassemble a chain.

4. Training and test corpora

Since we do not aim to test a particular text type, we want to use widest possible selection of texts. For this reason, short texts or fragments of texts in a variety of genres are included in the test corpus: news, scientific articles, blog posts and fiction. All texts are taken from Russian OpenCorpus. This corpus of texts in Russian has morphological, syntactical and semantic layers and is freely and fully accessible to researchers[1]. For test corpus about 200 texts are used, to the total of 500 texts. Number of texts used for train and test data can vary depending upon an experiment. 150 texts from different sections are used as the Golden Standard to check the model accuracy.

5. Measures

The performance of the model is evaluated by accuracy, precision and recall. Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations to the all observations in actual class. The latter metric is the most informative for our task.

To evaluate the performance of chains we used three measures for the coreference track evaluation: MUC-6, B3 and CEAF. They are normally used in anaphora and coreference resolution shared tasks. Precision, recall and F-measure are computed for each metric. More detailed information on this metrics can be seen in Toldova et al. RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian [10].

6.1. Baseline model

To perform first experiment no specific linguistic features were used to classify the pairs of mentions. We used the following features: distance between two mentions, relative distance (fraction of the distance and the number of symbols in the text), occurrence of mention 1 in mention 2, occurrence of mention 2 in mention 1, equality of mentions. A benefit of using gradient boosting is that after the boosted trees are constructed, it is relatively straightforward to retrieve importance scores for each feature. This importance is calculated explicitly for each feature in the dataset, allowing them to be ranked and compared to each other. Importance is calculated for a single decision tree by the amount that each feature split point improves the performance measure, weighted by the number of observations the node is responsible for. The feature importances are then averaged across all of the decision trees within the model. The most important feature is distance.

[1] http://opencorpora.org/

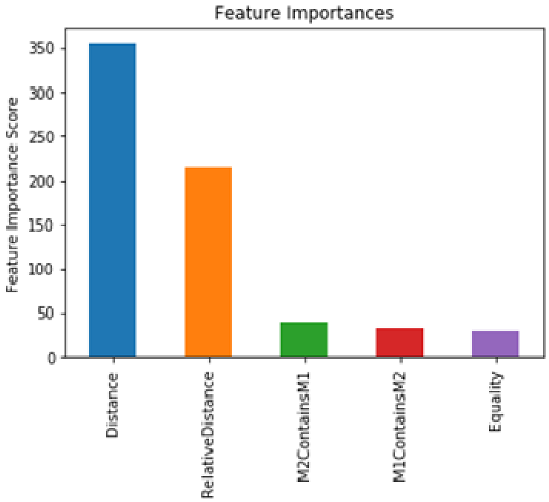

Fig.1. Feature importance

Dataset consists of 258509 pairs of mentions. For test data model showed 91% accuracy. Recall score was 23%, precision score was 43%. As measures for RU-EVAL-2019: evaluating anaphora and coreference resolution for Russian, we have the following metrics (see Fig. 2). To obtain the metrics we tested 180 documents, computed the three metrics and averaged them.

	Recall	Precision	F-measure
MUC-6	34.07	81.57	48.07
B-cube	21.95	82.72	34.7
CEAF-e	33.39	45.4	38.48

Fig. 2. Chains metrics

To analyze the errors we considered 100 random errors. These errors were analysed and it was concluded that the number of errors could be reduced by adding morphological features to the model. The largest percentage of errors refered to ‘0’ labels for a pronoun and its antecedent (40% of all errors). The same problem was in classification as non-coreferent two pronouns that had the same antecedent (15% of all errors). Such errors could be fixed by adding morphological features that dealt with parts of speech and faces of pronouns. The errors connected to labeling the same mentions in different cases as non-coreferent (13% of all errors) could be fixed by adding lemma features.

6.2. Experiment based on morphological and semantic features

We performed 3 experiments: experiment based on pair morphological features, experiment based on individual morphological features and experiment based on individual morphological and semantic features. The features set was added or changed according to error analysis of the previous experiment. The description of the final experiment can be found below.

To address the main error, we decided to add the following morphological pair features: lemma comparison, gender comparison, number comparison, animation category comparison. Moreover, we added the following noun and pronoun feature in the previous feature set: an antecedent is a noun (a left mention), an anaphor is a noun (a right mention), an antecedent is a pronoun, an anaphor is a pronoun, antecedent is a number, anaphor is a number, an antecedent is an adjective, an anaphor is an adjective. Defining adjective as either an antecedent or an anaphor could seem

inappropriate but we supposed that these features gave a system an opportunity to deal with mentions where the first word was an adjective. Moreover, we decided to use such tags from corpus morphological layers as a geographical place, an organization and a name in order to decrease the number of errors connected to classifying a pronoun and its antecedent as non-coreferent. Furthermore, we made an attempt to let the system take into account the fact that mentions could contain more than 1 word, therefore we added the following features: if an antecedent contains more than 3 words, if an anaphor contains more than 3 words and lemmas number that is equal in two mentions. Relying on the incorrect classification of synonyms, the following semantic features were added: distance between averaged semantic vectors of each mention in a pair, distance between maximum semantic vectors of each mention in a pair. These semantic features are basic semantic features for coreference resolution systems [12]. The current semantic features are based on semantic classes of words, since the semantic layer of the corpus consists of semantic embeddings of classes. A word can be referred to one class. There are 109903 semantic classes in the layer.
Accuracy score was 92.2. Recall and precision scores can be seen on Fig.3. [2]

Recall	Precision
43%	81%

Fig. 3. Model measures

The metrics for chains can be seen in Fig.4.

	Recall	Precision	F-measure
MUC-6	72.51	87.08	78.03
B-cube	71.36	71.14	71.02
CEAF-e	61.03	68.98	65

Fig. 4. Chains metrics

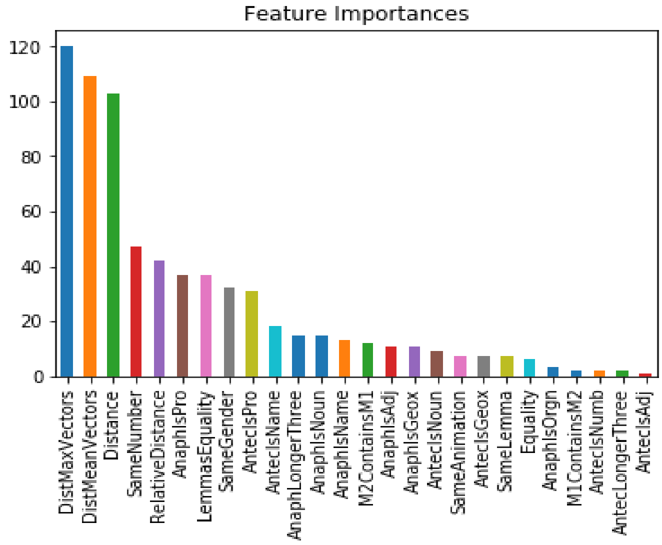

Fig. 5. Feature Importance

The both semantic features appeared to have the highest scores according to Feature Importance analysis. Distance between two mentions remained the main simple feature. Number comparison

[2] Jupyter notebook in html format with final experiment can be found here https://github.com/JuliaSilver/classifier

and gender comparison are the most influential among the morphological features. ‘if an anaphor (a right mention) is a pronoun’ is the most influential individual morphological feature. The least influential features is ‘if an antecedent is an adjective’ and ‘if an antecedent is a number’ but it can be connected to rarity of mentions with this label (see Fig.5)

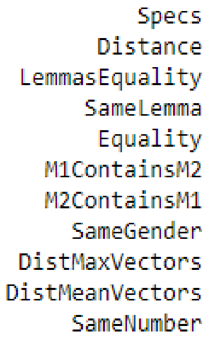

Fig. 6. Univariable analysis

According to Univariable analysis distance feature and lemma number comparison have the strongest relation with model output variable.

7. Conclusion and future plans

In this paper, we used a pairwise binary classification model as an approach to coreference resolution task for Russian to test morphological feature impact on the model accuracy. After computing metrics means for each experiment, we can conclude that morphological features addition raised the model accuracy about 15%. Semantic morphological features addition raised it 11.69% (see Fig. 7). The most influential features appeared to be semantic features, while distance features has the strongest relation with output variable.

	Model f-measure mean	Chains f-measure mean
1 experiment based on simple features	29.72%	40.42%
2 experiment based on pair morphological features	41.34%	63.77%
3 experiment based on individual morphological features	44.48%	66.52%
4 experiment based on individual morphological and semantic features	56.17%	71.35%

Fig. 7 Metrics means

After comparing the metrics change, we can conclude that the model accuracy and chains accuracy have been changing unevenly. That means that chain-reassembling algorithm impairs the model performance, therefore the chains accuracy is lower than model accuracy is. In our next experiments, we will reconsider chain-reassembling algorithm in order to deal with chains accuracy degradation. We aim to use other traditional chain-reassembling algorithms such as Closest antecedent algorithm and Best antecedent algorithm [12]. Moreover, we will treat the problem of chains reassembling as clusterization problem.

References

1) Bengtson E., Roth D., Understanding the value of features for coreference resolution, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, pp. 294-303, (2008).
2) Björkelund A., Farkas R., Data-driven Multilingual Coreference Resolution using Resolver Stacking, Joint Conference on EMNLP and CoNLL-Shared Task, Jeju Island, Korea, pp 49-55, (2012).
3) Budnikov E., Zvereva D., Maksimova D., Ru-Eval-2019: Evaluating Anaphora and Coreference Resolution for Russian, forum RU-EVAL-2019, Moscow, (2019).
4) Chen T., Guestrin C., Xgboost: A scalable tree boosting system, Proceeding KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, pp 785-794, (2016).
5) Clark K., Manning C., Entity-Centric Coreference Resolution with Model Stacking, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, pp 1405-1415, (2015).
6) Ng V., Cardie C., Improving machine learning approaches to coreference resolution, Proceedings of the 40th annual meeting on association for computational linguistics, Philadelphia, Pennsylvania, pp 104-111, (2002).
7) Ng V., Machine Learning for Entity Coreference Resolution: A Retrospective Look at Two Decades of Research, AAAI, San Francisco, pp 4877-4884, (2017).
8) Rahman A., Ng V., Supervised models for coreference resolution, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, pp 968–977, (2009).
9) Toldova S., Lyashevskaya O., Bonch-Osmolovskaya A., Ionov M., Evaluation for morphologically rich language: Russian NLP School of Linguistics, National Research University “Higher School of Economics”, Department of Theoretical and Applied Linguistics, Moscow State University, Moscow, pp 300-306, (2015).
10) Toldova S., Roytberg A., Ladygina A., Vasilyeva M., Azerkovich I., Kurzukov M., Sim G., Gorshkov D., Nedoluzhko A., Grishina Y. RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian, Moscow, (2014).
11) Yang X., Su J., Lang J., Tan C., Liu T., Li S., An entity-mention model for coreference resolution with inductive logic programming. Proceedings of Acl-08: Hlt, Columbus, Ohio, USA, pp 843–851, (2008)
12) Zheng J., Chapman W.R., Crowley S.G., Savova K., Coreference resolution: a review of general methodologies and applications in the clinical domain. Journal of Biomedical Informatics, Volume 44, Issue 6, pp 1113-1122, (2011)